Science of the Total Environment 409 (2011) 1559–1571

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM_{10} in Athens and Helsinki

A. Vlachogianni [a], P. Kassomenos [a,*], Ari Karppinen [b], S. Karakitsios [a], Jaakko Kukkonen [b]

[a] *Department of Physics, Laboratory of Meteorology, University of Ioannina, Greece*
[b] *Finnish Meteorological Institute (FMI), Air Quality Research, Finland*

ARTICLE INFO

Article history:
Received 5 February 2010
Received in revised form 30 August 2010
Accepted 30 December 2010
Available online 28 January 2011

Keywords:
PM_{10}
NO_x
Multiple regression analysis
Air quality forecasting
Model evaluation

ABSTRACT

Forecasting models based on stepwise multiple linear regression (MLR) have been developed for Athens and Helsinki. The predictor variables were the hourly concentrations of pollutants (NO, NO_2, NO_x, CO, O_3, $PM_{2.5}$ and PM_{10}) and the meteorological variables (ambient temperature, wind speed/direction, and relative humidity) and in case of Helsinki also Monin-Obukhov length and mixing height of the present day. The variables to be forecasted are the maximum hourly concentrations of PM_{10} and NO_x, and the daily average PM_{10} concentrations of the next day. The meteorological pre-processing model MPP-FMI was used for computing the Monin-Obukhov length and the mixing height. The limitations of such statistical models include the persistence of both the meteorological and air quality situation; the model cannot account for rapid changes (on a temporal scale of hours or less than a day) that are commonly associated, e.g., with meteorological fronts, or episodes of a long-range transport origin. We have selected the input data for the model from one urban background and one urban traffic station both in Athens and Helsinki, in 2005. We have used various statistical evaluation parameters to analyze the performance of the models, and inter-compared the performance of the predictions for both cities. Forecasts from the MLR model were also compared to those from an Artificial Neural Network model (ANN) to investigate, if there are substantial gains that might justify the additional computational effort. The best predictor variables for both cities were the concentrations of NO_x and PM_{10} during the evening hours as well as wind speed, and the Monin-Obukhov length. In Athens, the index of agreement (IA) for NO_x ranged from 0.77 to 0.84 and from 0.69 to 0.72, in the warm and cold periods of the year. In Helsinki, the corresponding values of IA ranged from 0.32 to 0.82 and from 0.67 to 0.86 for the warm and cold periods. In case of Helsinki the model accuracy was expectedly better on the average, when Monin-Obukhov length and mixing height were included as predictor variables. The models provide better forecasts of the daily average concentration, compared with the maximum hourly concentration for PM_{10}. The results derived by the ANN model where only slightly better than the ones derived by the MLR methodology. The results therefore suggest that the MLR methodology is a useful and fairly accurate tool for regulatory purposes.

1. Introduction

Pollution caused by particulate matter and nitrogen oxides is an issue of increasing public concern, due to its recognized adverse effects on human health. Numerous epidemiological studies have established the associations between the above-mentioned pollutants and daily excess in mortality (e.g., Dockery and Pope, 1994) and morbidity (e.g., Kassomenos et al., 2008). European Union (EU) has therefore established air quality standards for PM_{10}, involving an annual mean limit value of 40 $\mu g\ m^{-3}$ and a 24-hourly concentration limit (50 $\mu g\ m^{-3}$) that is not to be exceeded more than a specified number of times in a year. The air quality standards for NO_2 state that the maximum daily value must not exceed 200 $\mu g\ m^{-3}$ for more than 18 times a year.

High NO_x and PM_{10} concentrations are commonly measured in most European cities. The causes of these high values can be (a) local pollution sources, such as intensive traffic and small-scale combustion, (b) natural sources of particles (e.g., dust, sea salt and wild-land fires) (c) inefficient local atmospheric dispersion conditions (e.g., calm conditions, temperature inversions, etc.), and (d) synoptic weather conditions that favor the long-range transport of pollutants, especially particles (e.g., Vardoulakis and Kassomenos, 2008; Sofiev et al., 2009). An analysis of selected PM_{10} episodes in four European cities (Oslo, Helsinki, London and Milan) was carried out by Kukkonen et al. (2005). This article also presented a classification of European air pollution episodes. The best meteorological prediction variables were found to be the temporal evolution of the temperature inversions and atmospheric stability and, in some of the cases, wind speed.

* Corresponding author. Tel.: +30 2651008470; fax: +30 2651008671.
E-mail address: pkassom@uoi.gr (P. Kassomenos).

0048-9697/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2010.12.040

Numerous statistical models have been developed for forecasting urban air quality. The currently available statistical models are commonly neither applicable for predicting spatial concentration distributions in urban areas (except for some specific model types, such as the land-use regression models), nor for evaluating air pollution abatement scenarios for future years (Kukkonen et al., 2003). Deterministic statistical models are therefore more suitable for forecasting over extended areas such as major urban agglomerations and regional air quality. However, such models require precise input data, e.g., on the emissions and meteorological conditions.

Statistical models can establish relationships between input variables (predictors) and output variables (predictants), without detailing the causes of these relationships. Stochastic, multiple linear regression (Cordelino et al., 2001; Paschalidou et al., 2009) and neural network models (Kukkonen et al., 2003) have been established and shown to predict air pollutant concentrations with remarkable success. Multivariate statistical models for predicting daily concentrations of NO_x and PM_{10} have been previously applied in urban areas in several studies (Ping Shi and Harrison, 1997; Comrie and Diem, 1999; Abraham and Comrie, 2004; Chaloulakou et al., 2003, 2005; Kukkonen et al., 2003; Basurko et al., 2006; Lykoudis et al., 2008). On the other hand comparisons between multivariate statistical and neural network models indicated that neither type provide significant improvements of the forecasting ability (Comrie, 1997). Although in previous research statistical models have been used to forecast air pollution levels, these were not tested simultaneously in areas with different environmental and climatic conditions.

The main aims of this work were to construct accurate statistical models to forecast the concentrations of NO_x and PM_{10} based on meteorological and air quality data, and to critically evaluate the performance of such models. We have constructed linear predictive models, using stepwise multiple regression (MLR) for two European cities, Athens and Helsinki, and in addition applied Artificial Neural Network (ANN) models. The main reason for using the ANN models was to get a better insight on the usefulness of the MLR methods, i.e., to find out whether significantly more accurate results could be obtained with the substantially more complex statistical methods.

The two selected cities have substantially different characteristics. They are located in different climate zones (Mediterranean and sub-arctic), in different terrain (a basin surrounded by mountains, and fairly flat terrain), and have clearly different populations (4 and 1 million) and population densities. This study therefore provides insight on how well such statistical models could perform in substantially different urban environments and climatic conditions. Moreover significant parameters are identified not only providing a first means for predicting pollution levels, but also facilitating the development of more sophisticated tools for assisting environmental management and planning of future activities in the two areas.

2. Data and methods

2.1. Description of the two cities

Athens (37°58′N and 23°43′E) is a major urban agglomeration, with almost 4 million inhabitants and severe air pollution problems. It has been found out that insufficient city planning and the topographical features of the region surrounding the city enhance these problems. Athens is located in a basin having an extent of almost 450 km^2. The direction of the main axis of the basin is SSW–NNE. There are high mountains on three sides and the sea on one side of the basin.

The climate of Athens is Mediterranean; characteristic features include hot and dry summers and mild winters. The prevailing winds are from the Northern directions. The geography of the area does not therefore favor an efficient dispersion of air pollutants (Ziomas et al., 1995). Most of the local emissions of NO_x and PM_{10} originate from vehicular transport and industry. Natural sources, such as dust transportation from Sahara desert (Athanassiadou et al., 2006) and wild-land fires (Liu et al., 2009) play an important role in the determination of PM levels in Athens, especially during the warm period. During the cold season of the year, there are periods with very light winds and temperature inversions that do not have the forcing to clean the Athens basin (Kassomenos et al., 1999). There is also a period of consecutive days called “Halkyon days” in winter, with prevailing sunny and calm conditions that are responsible for the accumulation of air pollutants in the basin.

Helsinki Metropolitan Area (60°10′N and 25°0′E) and its surrounding region are situated on a fairly flat coastal area by the Baltic Sea at the latitude of 60°N (Kukkonen et al., 2005); the geographic extent is approximately 743 km^2 and the population is about 1 million inhabitants. Because of the warming effect of the Gulf Stream and the prevailing global circulation, the climate is relatively milder, compared to many other areas in the same latitudes (Kukkonen et al., 2005). The concentrations of NO_x and PM_{10} at street level are dominated by combustion, non-combustion and suspension emissions originating from vehicular traffic and long-range transport (in case of PM).

2.2. Air pollution and meteorological data

Hourly average concentrations of NO, NO_2, NO_x, CO, O_3, $PM_{2.5}$ and PM_{10} in 2005 were obtained from four stations, two for each city. The stations selected represent urban traffic and urban background environments. Specifically, for Athens, we used the station of Marousi (classification urban traffic, UT), which is located in the vicinity of an avenue with very high traffic during all days during all seasons and the station of Zografou (urban background, UB), which is located in the center of Athens, in an area that is not in a vicinity of any significant road.

For Helsinki we used and the stations of Vallila (UT) and Kallio (UB). The station of Vallila is situated in a park at a distance of 14 m from the edge of the Hämeentie road. The average weekday traffic volume of Hämeentie was 13,000 vehicles/day in 2001. The heights of the buildings in the vicinity of the station, at the other side of the Hämeentie road and surrounding the park, range from 10 to 15 m. The Hämeentie road is fairly wide; there are four lanes for cars and additionally two lanes for trams. The station of Kallio is located at the edge of a sports ground. The busiest streets in the vicinity of the station are Helsinginkatu at a distance of 80 m and Sturenkatu at a distance of 300 m. The average weekday traffic volume of Helsinginkatu was 7800 vehicles/day in 2001.

The meteorological parameters for Athens have been measured at the station of the National Observatory of Athens (NOA), located on the top of a hill (at a height of 107 m above the sea level), and have been found to be representative of the central urban area of the city (Kassomenos et al., 1998).

For Helsinki, the meteorological data were processed by the meteorological pre-processing model MPP-FMI (Karppinen et al., 2000a). The MPP-FMI model utilises meteorological synoptic and sounding observations, and its output consists of estimates of the relevant atmospheric turbulence parameters (the Monin-Obukhov length scale, the friction velocity and the convective velocity scale) and the boundary layer height. The meteorological data from the stations of Helsinki-Vantaa (about 15 km north of central Helsinki), Helsinki-Isosaari (an island about 20 km south of central Helsinki) and Jokioinen observatory (sounding data, about 115 km northwest from Helsinki) were taken into account. Unfortunately, it was not possible to use the MPP-FMI model in Athens, due to missing input data.

Using modeled instead of measured meteorological data for the location of central Helsinki was preferred, as it has previously been shown to be the best representative dataset with regard to this urban area. This dataset also contains relevant derived turbulence and boundary layer parameters, such as the Monin-Obukhov length and the mixing height (Karppinen et al., 2000a). The meteorological and air quality monitoring stations for both cities are presented in Fig. 1 a–b.

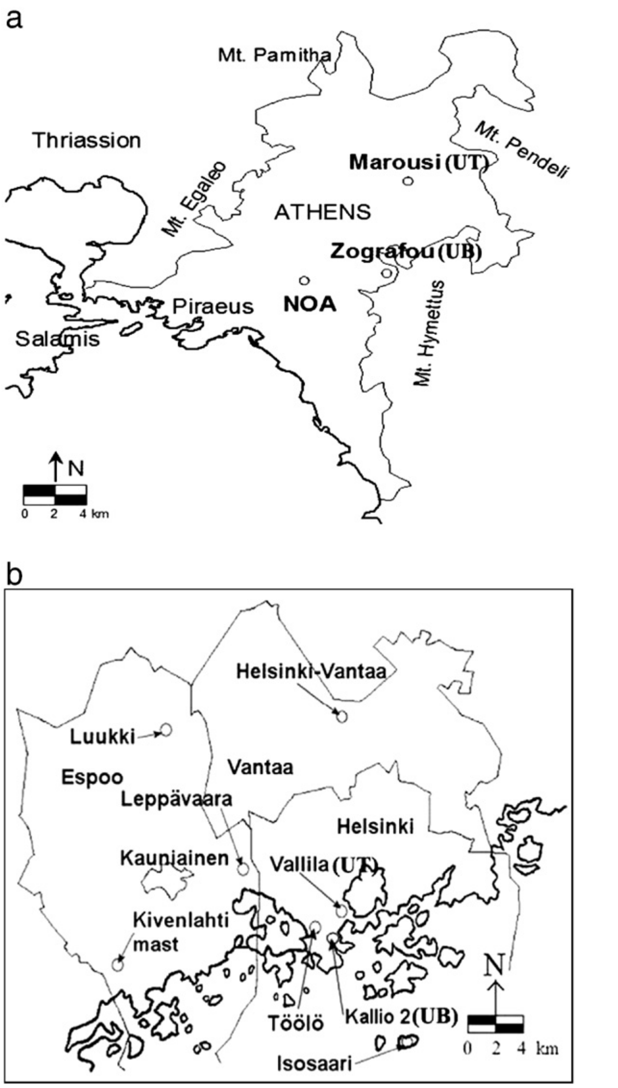

Fig. 1. Location of the air quality monitoring stations in (a) Athens and (b) Helsinki. The abbreviations UT and UB refer to air quality stations in urban traffic and urban background environments, respectively. The sites of Helsinki-Vantaa, Isosaari and Kivenlahti mast are meteorological stations; Helsinki, Vantaa, Espoo and Kauniainen are cities.

For Athens, to avoid the discontinuity that would be caused by using wind direction (WD), expressed in degrees, we used the Wind Direction Index (WDI) (Lalas et al., 1982; Ordieres et al., 2004):

$$\mathrm{WDI} = 1 + \sin\left(\mathrm{WD} + \frac{\pi}{4}\right). \tag{1}$$

The quality of the datasets was first examined. We did not use data rows, for which one or more of the model input variables were missing; these were categorized as missing data. The fraction of missing data ranged from 3.5% to 16.0% for Athens and from 1.1% to 4.1% for Helsinki, respectively (depending on the station). For instance, in case of PM_{10}, the entry/missing data are for the cold period at Marousi 157/25, Zografou 169/13, Kallio 180/2 and Vallila 176/6. For the warm period the respective data are for Marousi 175/8, Zografou 173/10, Kallio 181/2 and Vallila 174/9.

We separated the datasets into two periods: a) the so-called warm period, i.e., the half yearly period from April to September, and b) the cold period, i.e., the half yearly period from October to March; excluding January 1st. We randomly selected 25% of the final data of each period to be used for the independent evaluation of the model, and used the remaining 75% of the data for the training of the model.

2.3. Multiple linear regression analysis

Multiple linear regression analysis has been used in the field of air pollution by several authors (Comrie and Diem, 1999; Barcenas et al., 2005; Basurko et al., 2006; Kolehmainen et al., 2001). The general form of multiple linear regression analysis is the following:

$$Y_i = \beta_0 + \beta_1 X_{1\iota} + \beta_2 X_{2\iota} + \ \ldots + \beta_p X_{pi} + \varepsilon_i \tag{2}$$

where Y is the dependent variable (predictant), β_0 is a constant coefficient, β_1, β_2,…β_p are the regression coefficients of the independent variables X_1, X_2,…X_p (predictors) and ε is the residual error (difference between observations and predicted values).

The assumptions required to apply multiple linear regression are: (i) the predictor variables must be independent and (ii) the residual errors ε_i must be independent and they have to be normally distributed, with a vanishing mean and a constant variance (σ^2). The least squares method is a simple and widely preferred procedure for the estimation of the parameters β_1, β_2,…β_p.

We have employed a stepwise regression procedure to select the independent variables that would result in the best possible model, while at the same time ensuring statistical significance of the results. In this method the best predictor variables, according to some statistical criterion, are entered into the prediction equation, one after the other in successive steps, until no other predictor variable meets the criterion. At each step the variables already entered into the equation are checked; i.e., if they still satisfy the statistical criterion.

In this specific case, the predictors entering the model were selected to be the ones with the largest partial correlation with the dependent variable. This procedure controls the selection of the predictors in the regression model. In addition, the partial regression coefficient of a predictor must be significant at the 0.05 level, and at least 0.01% of its variance has to be independent of the other predictor variables (tolerance value) in order to be selected (SPSS Inc., 2000).

In all cases, the distribution of residuals (e.g., the differences between the observed and computed values of the dependent variable) is normal. Normal probability plot is used to determine whether the distribution of residuals matches a specified, in this case normal, distribution.

In this study, regression models were developed for the maximum hourly PM_{10} and NO_x concentrations, as well as for the daily average PM_{10} concentrations of the next day, separately for the cold and warm periods of the year. We used as independent variables the hourly values of the various meteorological parameters and pollutants of the present day. For instance, the independent variables representing PM_{10} comprised of H0PM_{10}, H1PM_{10},…,H23PM_{10}, where the notations H0, H1,…H24 refer to the hour of the present day.

We performed the regression model computations for Helsinki in two ways using, apart from the pollutant concentrations, (i) only the measured meteorological variables, namely temperature, relative humidity, and wind speed and direction (denoted as Case A) and (ii) using also the selected turbulence and boundary layer parameters, namely the Monin-Obukhov length and the mixing height, computed using the MPP-FMI model (Case B).

In order to examine the results obtained by the MLR, we compared them with already available results obtained by an Artificial Neural Network (ANN) model. This ANN model (multi-layer perceptron) uses the BFGS quasi-Newton algorithm (Dennis and Schnabel, 1983) for training purposes. It uses 65% of the whole dataset as a training set, and 10% as a testing set, in the training procedure. The trained model was subsequently evaluated by an independent dataset, characterized as a validation test; this data consisted of 25% of the overall dataset. Such an evaluation against independent data was performed for both the MLR and the ANN models. The comparison of the results of the two models (MLR and ANN) will be presented in the evaluation of the model performance section, by means of Pearson correlation coefficients.

2.4. Evaluation of the methods

To evaluate the models we used a data processing program, which is included in the Model Validation Kit (Basurko et al., 2006; Niska et al., 2004, 2005). This includes several model performance measures such as the R (Pearson correlation coefficient), NMSE (normalized mean square error), FA_2 (factor of two), FB (fractional bias) and FV (fractional variance) (Basurko et al., 2006). We also used other statistical model performance measures such as the MBE (mean bias error), MAE (mean absolute error), RMSE (root mean square error), and IA (index of agreement). These have been discussed, e.g. by Willmott (1981) and Karppinen et al. (2000b).

Letting O_i be the observed and P_i the forecasted values, $\sigma_{C_o}$and σ_{C_p} are the respective standard deviations and $\overline{O}$ and $\overline{P}$ are the respective means. The definitions of the statistical measures of the goodness of fit used herein are the following:

$$\text{(i)MBE} = \overline{P} - \overline{P} \tag{3}$$

$$\text{(ii)MAE} = \frac{\sum_i |O_i - P_i|}{N} \tag{4}$$

$$\text{(iii)RMSE} = \sqrt{\frac{\sum_i^N (O_i - P_i)^2}{N}} \tag{5}$$

$$\text{(iv)IA} = 1 - \frac{\sum_i^N (O_i - P_i)^2}{\sum_i^N (|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \tag{6}$$

$$\text{(v)}R = \frac{\overline{(O_i - \overline{O_i}).(P_i - \overline{P_i})}}{\sigma_{P_i}.\sigma_{P_o}} \tag{7}$$

$$\text{(vi)NMSE} = \frac{\overline{(P_i - O_i)^2}}{\overline{O_i}\,\overline{P_i}} \tag{8}$$

$$\text{(vii)FA}_2 = \frac{\sum_\iota k_i}{N}, \quad k_i = \begin{cases} 1, \text{ ..if...} 0.5 \le \dfrac{P_i}{O_i} \le 2 \\ 0, \text{else} \end{cases} \tag{9}$$

$$\text{(viii)FV} = \frac{\sigma_{C_o} - \sigma_{C_p}}{\frac{1}{2}\left(\sigma_{C_o} + \sigma_{C_p}\right)} \tag{10}$$

$$\text{(ix)FB} = \frac{2\left(\overline{P}_i - \overline{O}_i\right)}{\overline{P}_i - \overline{O}_i}. \tag{11}$$

Mean Biased Error (MBE) indicates the degree of overprediction (MBE>0) or underprediction (MBE<0) of the observed concentrations (Comrie, 1997). The observed σ_{C_o} and modeled σ_{C_p} concentration standard deviations quantify the amount of the observed variance captured by the model. The two commonly reported measures of residual error, MAE and RMSE, summarise the difference between the observed and modeled concentrations. Due to the power term in the RMSE calculation, it is more sensitive to extreme values than MAE (Gardner and Dorling, 2000).

IA is a relative and bounded measure that allows for cross-comparisons between models, and is limited to the range of 0–1 (Lu, 2003).

The Pearson correlation coefficient between P_i and O_i, R, quantifies the overall performance of the model (R), the coefficient of determination (R^2) and IA are measures that indicate similarity between the model tendency and the observed one (Pastor-Barcenas et al., 2005).

The Normalized Mean Square Error, NMSE, is an estimator of the overall deviations between observed and predicted concentrations. NMSE generally shows the most striking differences among the models. If a model has a low NMSE, then it performs well. On the other hand, high NMSE values do not necessarily mean that the model is completely wrong. The differences of peak values have a higher weight on NMSE than the differences of other values (Willmott,1981).

The factor of two, FA_2, gives the percentage of cases, in which the values of the ratio O_i/P_i are in the range [0.5, 2] (Basurko et al., 2006). Finally, the Fractional Variance, FV, is a normalized measure that allows the comparison of differences between the predicted variance and the observed variance. A model with FV = 0 is a model whose variance is equal to the variance of the observed values (Basurko et al., 2006).

3. Results and discussion

We have first presented and interpreted physically the temporal variations of the measured concentrations at the stations considered, as part of the QA/QC of the data, and for understanding better the behavior of the datasets that will be used as model input.

3.1. The diurnal and seasonal variation of the measured concentrations

3.1.1. Athens

The main sources of PM_{10} and NO_x in Athens are the emissions of 2.5 million vehicles registered in the area, whereas domestic heating in winter and light industrial activities in the southwestern parts of the city are secondary sources. The most important sources of PM_{10} are local vehicular traffic, particulate matter from wild-land fires (Liu et al., 2009), desert dust and sea salt.

The diurnal and seasonal variations of measured NO_x concentrations in 2005 for cold and warm seasons are presented in Fig. 2 a–b for both stations (urban traffic and urban background) in case of Athens. The diurnal variation was presented as an average over the warm and cold periods, respectively, including both weekdays and weekends. The diurnal variation in the two stations reproduces the concentration peaks in the morning and in the evening: these were found previously by Kassomenos (2005). The evening peak occurs late (the maximum value at 23:00), compared with the corresponding peak in most major cities in Central and Northern Europe. The main explanation is associated to the social and economic lifestyles, e.g., after 19:00 hour people are returning home after the commercial stores close for the night, while some other people are going out for recreational purposes.

In case of Athens, there is no seasonal variability of the NO_x concentrations at the urban background station. On the contrary, at the urban traffic station, there is an intensive seasonal variability, with higher concentrations in winter and lower ones in summer. The latter is due to both the reduction of the traffic flows during the summer period (Kassomenos, 2005) and the increase of the photochemical activity that reduces NO_x and produces O_3, since the area experiences intensive solar insolation (Paschalidou and Kassomenos, 2004).

The diurnal variability of PM_{10} (Fig. 2 c) reflects the activities of the people working around the sites; there is a peak of concentrations during the rush hours at the urban traffic station and an almost constant behavior after 11.00 at the urban background station. The seasonal variability of PM_{10} at the urban background station (Fig. 2 d) reveals spring and autumn peaks, probably due to the contribution of natural sources (Borge et al., 2007). The seasonal variability at the urban traffic station is not clear; the same result has also been observed during 2001–2003 (Vardoulakis and Kassomenos, 2008).

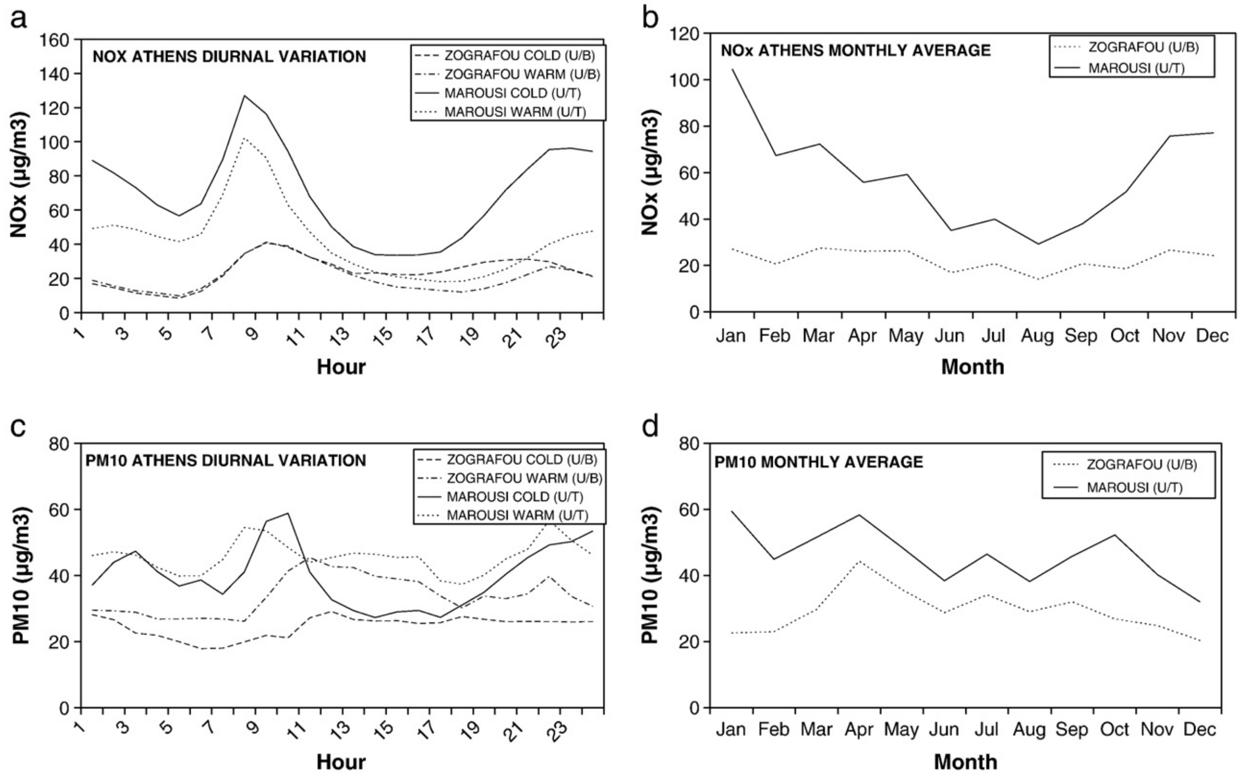

Fig. 2. Diurnal and seasonal variations of NO_x (panels a and b) and PM_{10} (c and d) concentrations in Athens in 2005.

3.1.2. Helsinki

The concentrations of NO_x and PM_{10} at street level are dominated by vehicular emissions and in case of PM_{10}, long-range transport. There are occasional episodes of high PM_{10} concentrations caused by long-range transport of anthropogenic pollution from Central and Eastern Europe, and wild-land fires originating from Russia, Baltic countries and Central Eastern Europe.

We have estimated the diurnal variations of the NO_x concentrations in Helsinki, at the urban traffic and background stations (Fig. 3 a). The NO_x concentrations are continuously higher at both stations from

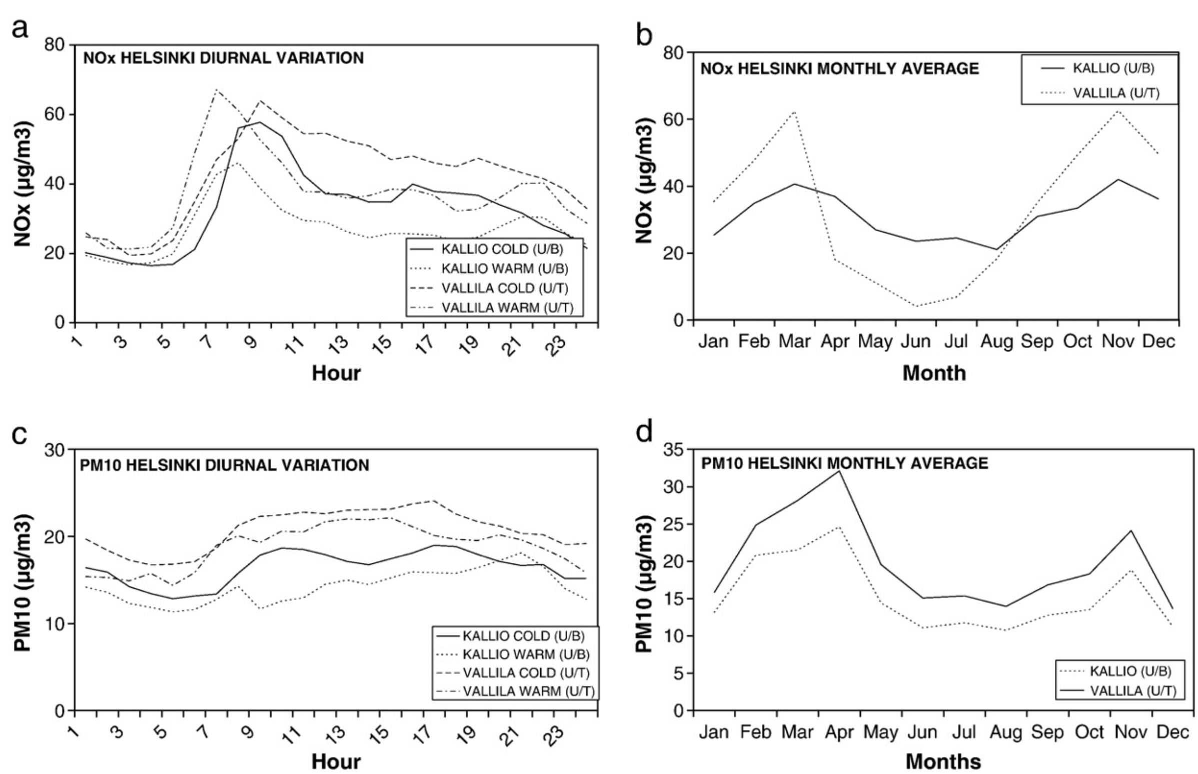

Fig. 3. Diurnal and seasonal variations of NO_x (panels a and b) and PM_{10} (c and d) concentrations in Helsinki in 2005.

the morning rush hour (at 8:00) until the evening hours (approximately from 19:00 to 22:00). However, no distinct maxima during rush hours, such as those in Athens, are detected. This behavior is probably caused by the less severely congested traffic in Helsinki, compared with Athens.

In case of the seasonal variations, higher NO_x concentrations occur in spring and autumn, and the minimum occurs in summer. This variation is associated with the seasonal variation of the relevant meteorological dispersion conditions; such as the more common occurrence of stable and extremely stable low wind speed situations during the winter half-year (Kukkonen et al., 2005) (Fig. 3 d).

The diurnal and seasonal variations of the PM_{10} concentrations are presented in Fig. 3 c–d for Helsinki. The PM_{10} concentrations present a similar behavior with the NO_x concentrations (Aarnio et al., 2007).

3.2. Statistical analysis of the observed and predicted values of the dependent variables

We first examine the statistical significance of the correlations between the dependent variables of the models and the various predictors, as this is crucial in terms of the inclusion of these predictors in the final stepwise regression models. Second, the model performance is evaluated using an independent dataset that has not been used in the setting of the values of these constants.

Descriptive statistics for both the observed values and for those predicted by the regression model in case of Athens are presented in Tables 1a–1b. The highest hourly averaged values in a day for both pollutants considered have been presented in Table 1a, and the daily average values for PM_{10} have been presented in Table 1b. The corresponding statistical values in case of Helsinki have been presented in Tables 2a–2b.

The coefficient of determination squared, adjusted for the number of regressed points (R^2) (Table 1a) of the highest hourly averaged values in a day for both the warm and cold periods in Athens ranged from 0.37 to 0.65. The model forecasted with a relatively higher accuracy the NO_x concentrations during the warm period and the PM_{10} concentrations during the cold period, compared with the corresponding forecasts for the other period.

The coefficient of determination squared of the daily average concentrations ranged from 0.50 to 0.86, for both the warm and cold periods in Athens, for PM_{10} (Table 1b). As expected, the model ability is better for the forecasts of the daily average values, compared with those for the highest hourly values in a day. Clearly, this is caused by the fact that longer integrating times tend to smooth out shorter-term fluctuations that are more difficult to predict reliably. The model forecasted the daily average values with a higher accuracy during the cold period, compared with the warm period.

In case of Helsinki, the model computations were performed using two modeling options denoted by A and B. In Case A, the meteorological data that were based directly on measured quantities (including only a spatial interpolation), i.e., ambient temperature, wind speed and direction, and relative humidity, were used as input variables of the model. In Case B, additionally Monin-Obukhov length and mixing height were used as input variables of the model; these were computed based on the boundary layer scaling contained in the MPP-FMI model.

The descriptive statistics of the measured and predicted values, and the regression coefficients of the measured and forecasted time series have been presented in Tables 2a–2b. As in the case of Athens, the model ability was on the average better for the forecasts of the daily average values, compared with those for the highest hourly values in a day. The model ability, as measured by the coefficient of determination, was in all cases except for one better or at least equal in Case B, compared with Case A. The inclusion of turbulence and boundary layer parameters as additional input variables therefore in most cases improves the forecasting capability. Clearly, this result is to be expected, as the turbulence strength and boundary layer height are key parameters for the forecasting of air quality, especially in episodic conditions (e.g., Kukkonen et al., 2005).

Table 1b
Statistical parameters of the predicted and observed daily average concentrations for PM_{10}, presented separately for the warm and cold periods, for Athens.

Daily average concentration of PM_{10}				
Statistical parameters	Zografou UB		Marousi UT	
	Predicted	Observed	Predicted	Observed
Mean ($\mu g\ m^{-3}$)				
Warm period	15	34	48	46
Cold period	46	33	42	40
Max ($\mu g\ m^{-3}$)				
Warm period	56	96	110	174
Cold period	450	464	107	82
St. dev. ($\mu g\ m^{-3}$)				
Warm period	15	16	19	26
Cold period	75	59	22	18
R^2				
Warm period	0.80		0.50	
Cold period	0.86		0.64	

UB = urban background station, UT = urban traffic station, St. dev. = standard deviation, R^2 = coefficient of determination squared.

Table 1a
Statistical parameters of the predicted and observed highest hourly concentrations in a day, for NO_x and PM_{10}, presented separately for the warm and cold periods, for Athens.

Statistical parameters	Highest hourly daily concentrations in a day of NO_x				Highest hourly concentrations in a day of PM_{10}			
	Zografou UB		Marousi UT		Zografou UB		Marousi UT	
	Predict.	Observ.	Predict.	Observ.	Predict.	Observ.	Predict.	Observ.
Mean ($\mu g\ m^{-3}$)								
Warm period	62	57	125	125	63	62	94	90
Cold period	67	62	213	206	54	52	102	107
Max ($\mu g\ m^{-3}$)								
Warm period	133	238	409	458	356	433	240	288
Cold period	145	205	661	847	459	484	222	474
St. dev. ($\mu g\ m^{-3}$)								
Warm period	29	37	76	89	68	46	33	33
Cold period	22	35	124	160	46	57	44	67
R^2								
Warm period	0.61		0.65		0.57		0.45	
Cold period	0.37		0.60		0.58		0.58	

UB = urban background station, UT = urban traffic station, St. dev. = standard deviation, R^2 = coefficient of determination squared.

Table 2a
Statistical parameters of the predicted and observed highest hourly concentrations in a day, for NO_x and PM_{10}, presented separately for the warm and cold periods, and for the two modeling options denoted by the letters A and B. In Case A, the meteorological data was based directly on measured quantities, and in Case B, additionally Monin-Obukhov length and mixing height were used as input variables of the model.

Statistical parameters	Highest hourly concentrations in a day of NO_x				Highest hourly concentrations in a day of PM_{10}			
	Kallio UB		Vallila UT		Kallio UB		Vallila UT	
	Predict.	Observ.	Predict.	Observ.	Predict.	Observ.	Predict.	Observ.
Mean (µg m^{-3})								
Warm period								
A	63	65	97	103	30	29	32	33
B	62	65	98	103	30	29	32	33
Cold period								
A	98	95	126	134	34	34	38	39
B	99	95	126	134	34	34	38	39
Max (µg m^{-3})								
Warm period								
A	172	298	335	493	112	194	106	172
B	286	298	441	493	112	194	106	172
Cold period								
A	283	1025	1093	1358	136	215	144	157
B	291	1025	1164	1358	136	215	142	157
St. dev. (µg m^{-3})								
Warm period								
A	25	47	36	80	16	20	16	22
B	35	47	59	80	16	20	16	22
Cold period								
A	68	113	128	167	21	26	22	27
B	59	113	134	167	21	26	25	27
R^2								
Warm period								
A	0.32		0.27		0.48		0.67	
B	0.53		0.61		0.48		0.70	
Cold period								
A	0.57		0.60		0.45		0.24	
B	0.68		0.70		0.55		0.30	

UB = urban background station, UT = urban traffic station, St. dev. = standard deviation, R^2 = coefficient of determination squared..

Tables 3 and 4a–4b present the predictors included by the multiple stepwise regression in the predicting equations for both the warm and cold periods, for Athens and Helsinki (Cases A and B). We have split the present day in two sub-periods, 1–12 h named as *m* and 13–24 h, named as *a*. So, if the value of a predictor was measured during the hours 1–12 of the present day, it is counted to column *m*, while if it was measured during the hours 13–24, it is counted to column *a* for each predictor. The best predictor variables and the correlation coefficients (R) between the best predictor and the dependent variables are presented in the same tables.

In case of Athens, the forecasted NO_x concentrations are strongly dependent on the NO_2, NO and CO concentrations measured on the present day, especially at the urban traffic (UT) station, and the effect of meteorology (especially wind speed and direction) is also important. For PM_{10}, both for forecasted hourly maximum and daily average values, PM_{10} concentrations measured on the present day are significant contributors.

For the prediction of average PM_{10} values of the next day, the local meteorology plays a smaller role than for the prediction of the hourly maximum concentrations of both NO_x and PM_{10}. This is caused by the

Table 2b
Statistical parameters of the predicted and observed values daily average concentrations for PM_{10}, presented separately for the warm and cold periods, and for the two modeling options denoted by the letters A and B, for Helsinki.

Statistical parameters	Daily average concentration of PM_{10}							
	Warm period				Cold period			
	Kallio UB		Vallila UT		Kallio UB		Vallila UT	
	Predict.	Observ.	Predict.	Observ.	Predict.	Observ.	Predict.	Observ.
Mean (µg m^{-3})								
Case A	15	14	19	19	17	17	21	21
Case B	15	14	19	19	17	17	21	21
Max (µg m^{-3})								
Case A	43	51	56	62	49	45	69	108
Case B	43	51	56	62	45	45	85	108
St. dev. (µg m^{-3})								
Case A	7	9	8	10	8	9	11	13
Case B	7	9	9	10	8	9	12	13
R^2								
Case A	0.66		0.72		0.58		0.63	
Case B	0.66		0.70		0.66		0.65	

UB = urban background station, UT = urban traffic station, St. dev. = standard deviation, R^2 = coefficient of determination squared.

Table 3
Predictors included by the multiple stepwise regression in the predicting equations for both the warm and cold period for Athens.

Dependent variable	NO_2 m[a]	NO_2 a[a]	NO_x m[a]	NO_x a[a]	NO m[a]	NO a[a]	CO m[a]	CO a[a]	O_3 m[a]	O_3 a[a]	PM_{10} m[a]	PM_{10} a[a]	WS m[a]	WS a[a]	WDI m[a]	WDI a[a]	T m[a]	T a[a]	RH m[a]	RH a[a]	BPR[b]	R[c]	R^2[d]	VIF[e]
Warm period																								
MAR NO_x	X	X			X	X		X			X		X						X		$H22NO_2$[f]	0.79	0.69	2.46
ZOG NO_x						X				X				X		X	X		X		$H20NO_2$[c]	0.65	0.58	2.91
MAR PM_{10}	X	X						X				X	X		X						H10WDI[f]	0.67	0.48	2.41
ZOG PM_{10}					X				X			X		X							$H16PM_{10}$[c]	0.67	0.58	2.11
MAR av. PM_{10}		X		X				X			X	X									$H6PM_{10}$[f]	0.80	0.72	2.36
ZOG av. PM_{10}		X			X				X	X	X	X									$H24PM_{10}$[f]	0.52	0.77	2.54
Cold period																								
MAR NO_x					X			X				X		X		X				X	H24CO[f]	0.81	0.75	2.46
ZOG NO_x	X				X	X									X				X		H24WS[f]	0.65	0.49	2.03
MAR PM_{10}	X			X										X		X					H24NOx[f]	0.86	0.80	2.30
ZOG PM_{10}												X	X	X	X				X	X	$H24PM_{10}$[f]	0.86	0.79	1.25
MAR av. PM_{10}				X				X				X									$H22PM_{10}$[f]	0.60	0.71	2.21
ZOG av. PM_{10}					X	X						X				X					H24NO[f]	0.74	0.87	2.62

The highest six rows of the table correspond to the warm period, and the lowest six rows to the cold period. The notation av. refers to the prediction of daily average values (the other values are highest hourly concentrations), WDI and WS are the wind direction index and wind speed.
[a] m refers to the hours 01–12; a to the hours 13–24.
[b] BPR = best predictor variable found using multiple regression analysis.
[c] R = correlation coefficient between the best predictor and the dependent variables.
[d] R^2 = coefficient of determination squared of the model.
[e] VIF = variance inflation factor (maximum value).
[f] Hxx = hourly averaged concentration on a specific hour of the day, e.g. $H20NO_2$ refers the NO_2 concentration at 8 p.m.

fact that especially the elevated daily averaged PM_{10} concentrations can commonly be caused by regionally or long-range transported pollution; such episodes are not sufficiently described by the locally measured meteorological conditions.

In most cases, the best predictor variables of both NO_x and PM_{10} are statistically most significantly associated to the NO_x and PM_{10} concentrations of the present day, especially during the evening hours of the present day. Regarding the meteorological parameters, wind speed and direction are most substantially associated to the dependent variables. However, there is also a substantial variation in the best predictor variables (i.e., these are different from case to case) for both NO_x and PM_{10}. These variations are partly due to the local urban characteristics of each station, such as the detailed distribution of pollutant sources in the vicinity of the stations.

For Helsinki, we examined separately Cases A (Table 4a) and B (Table 4b). In Case A, the independent parameters are the same as those for Athens, whereas in Case B, additionally Monin-Obukhov length and mixing height are included. In Case A, the measured concentrations of NO_x and PM_{10} from the present day are the main contributors, while the contribution of meteorological values is less significant than for Athens. For Case B, also the meteorological parameters, especially Monin-Obukhov length and mixing height contribute significantly.

Similarly to the case of Athens, in most cases, the best predictors are the corresponding concentrations of the selected two pollutants during the evening hours of the present day.

3.3. Evaluation of the model performance

We present the model evaluation against an independent dataset that was not used in setting the values of the model constants. This randomly selected independent dataset contains almost 25% of the total data. We have used various statistical indexes and skill measures, such as those included in the Model Validation Kit (Olesen, 1995). The

Table 4a
Predictors included by the multiple stepwise regression in the predicting equations for both the warm and cold periods for Helsinki, in Case A.

Dependent variable	NO_x m[a]	NO_x a[a]	NO_2 m[a]	NO_2 a[a]	NO m[a]	NO a[a]	O_3 m[a]	O_3 a[a]	SO_2 m[a]	SO_2 a[a]	PM_{10} m[a]	PM_{10} a[a]	$PM_{2.5}$ m[a]	$PM_{2.5}$ a[a]	WS m[a]	WS a[a]	T m[a]	T a[a]	RH m[a]	RH a[a]	BPR[b]	R[c]	R^2[d]	VIF[e]
Warm period																								
Val NO_x		x										x									$H24NO_x$[f]	0.80	0.69	2.43
Kall NO_x		x	x																		$H24NO_2$[f]	0.85	0.78	2.39
Val PM_{10}												x					x			x	$H19PM_{10}$[f]	0.80	0.72	3.01
Kall PM_{10}												x	x				x				$H22PM_{10}$[f]	0.60	0.56	1.99
Vall av. PM_{10}		x							x	x		x					x				$H20PM_{10}$[f]	0.72	0.81	2.94
Kall av. PM_{10}		x										x	x								$H24PM_{10}$[f]	0.79	0.67	2.21
Cold period																								
Vall NO_x											x	x			x						$H21PM_{10}$[f]	0.38	0.67	2.13
Kall NO_x		x					x													x	$H23NO_x$[f]	0.41	0.42	3.02
Vall PM_{10}							x					x		x							$H21PM_{10}$[f]	0.70	0.72	2.37
Kall PM_{10}		x	x									x									$H21PM_{10}$[f]	0.86	0.84	2.19
Vall av. PM_{10}		x	x									x									$H21PM_{10}$[f]	0.68	0.71	3.27
Kall av. PM_{10}			x									x		x							$H23PM_{10}$[f]	0.86	0.81	1.98

[a] m: refers to the hours 01–12; a to the hours 13–24.
[b] BPR: best predictor variable found using multiple regression analysis.
[c] R: correlation coefficient between the best predictor and the dependent variables.
[d] R^2 = coefficient of determination squared of the model.
[e] VIF = variance inflation factor (maximum value).
[f] Hxx: hourly averaged concentration on a specific hour of the day.

Table 4b
Predictors included by the multiple stepwise regression in the predicting equations for both the warm and cold periods for Helsinki, in Case B.

Dependent variable	NO_x m[a]	NO_x a[a]	NO_2 m[a]	NO_2 a[a]	O_3 m[a]	O_3 a[a]	SO_2 m[a]	SO_2 a[a]	PM_{10} m[a]	PM_{10} a[a]	$PM_{2.5}$ m[a]	$PM_{2.5}$ a[a]	T m[a]	T a[a]	RH m[a]	RH a[a]	INVL m[a]	INVL a[a]	MIX m[a]	MIX a[a]	BPR[b]	R[c]	R^2[d]	VIF[e]
Warm period																								
Val NO_x		x							x				x	x				x		x	$H24NO_x$[f]	0.29	0.74	2.22
Kall NO_x				x					x		x		x					x			H24INVL[f]	0.47	0.67	2.15
Val PM_{10}										x			x			x		x	x		$H19PM_{10}$[f]	0.76	0.70	2.03
Kall PM_{10}										x	x		x								$H22PM_{10}$[f]	0.60	0.48	1.17
Vall av. PM_{10}		x								x		x						x			$H20PM_{10}$[f]	0.68	0.73	2.09
Kall av. PM_{10}		x								x		x									$H24PM_{10}$[f]	0.79	0.74	2.36
Cold period																								
Vall NO_x		x							x							x					$H21NO_x$[f]	0.38	0.68	3.19
Kall NO_x		x				x										x		x	x		$H23NO_x$[f]	0.41	0.48	2.28
Vall PM_{10}	x						x			x					x						$H21PM_{10}$[f]	0.70	0.76	2.37
Kall PM_{10}		x		x	x					x	x								x		$H21PM_{10}$[f]	0.74	0.60	2.81
Vall $DAPM_{10}$		x								x							x		x		$H21PM_{10}$[f]	0.73	0.64	2.19
Kall $DAPM_{10}$					x					x							x		x		$H23PM_{10}$[f]	0.73	0.75	1.98

[a] m: refers to the hours 01–12; a to the hours 13–24.
[b] BPR: best predictor variable found using multiple regression analysis.
[c] R: correlation coefficient between the best predictor and the dependent variables.
[d] R^2 = coefficient of determination squared of the model.
[e] VIF = variance inflation factor (maximum value).
[f] Hxx: hourly averaged concentration on a specific hour of the day.

results of the application of these measures are presented in Table 5 for Athens and in Tables 6 and 7 for Helsinki for the MLR model.

For Athens, the model performance measures for the maximum hourly values in a day are in most cases better during the warm period, compared with the cold period. However, for the daily average concentrations of PM_{10}, the model has in most cases higher accuracy during the cold than warm period. The model forecasts with a higher accuracy in three cases from the four cases considered, when the dependent variable is daily average concentration of PM_{10}, compared with the corresponding forecasts for the maximum hourly concentration.

The model performance is in all cases better for NO_x, compared with the corresponding results for PM_{10}. The main reason is probably that the urban NO_x concentrations are to a large extent originated from local sources, while the urban PM_{10} concentrations are more commonly of a regional or long-range transported (LRT) origin (which are not explicitly taken into account in such a statistical model).

For Helsinki, the model has a better performance in most of Case B (i.e., using additionally pre-processed meteorological data as model input), compared with the corresponding Case A. In the warm period, the model performance is better for PM_{10}, compared with the corresponding cases for NO_x. However, in most cases in the cold period, the forecasts for NO_x are better than those for PM_{10}. For both warm and cold periods, the model performance is in most cases better for the forecasting of the daily average PM_{10} compared with the corresponding forecasts for the hourly maximum values in a day.

Table 5
Statistical model performance measures using an independent dataset, for Athens.

Stations	Pollut.	r[a]	NMSE[b]	FA_2[c]	FB[d]	FV[e]	MBE[f]	MAE[g]	RMSE[h] ($\mu g\ m^{-3}$)	IA[i]
Highest hourly concentrations in a day										
Warm period										
Marousi	NO_x	0.73	0.25	0.86	−0.03	0.16	2.97	44.83	55.36	0.84
	PM_{10}	0.73	0.15	0.93	−0.07	0.27	6.74	22.86	31.78	0.81
Zografou	NO_x	0.64	0.26	2.01	−0.1	−0.1	5.79	21.68	27.22	0.77
	PM_{10}	0.59	0.07	1.04	0.009	0.12	−0.57	21.37	27.22	0.77
Cold period										
Marousi	NO_x	0.69	0.93	0.55	0.35	0.43	−64.27	137.29	170.91	0.69
	PM_{10}	0.50	0.37	0.99	0.09	0.53	−10.99	47.65	67.33	0.62
Zografou	NO_x	0.60	0.23	0.91	−0.01	0.44	5.75	22.04	28.18	0.72
	PM_{10}	0.22	0.21	0.88	−0.15	0.21	7.15	16.52	21.31	0.40
Daily average concentrations of PM_{10}										
Warm period										
Marousi	PM_{10}	0.81	0.13	1.13	−0.02	0.28	1.10	6.79	12.99	0.86
Zografou	PM_{10}	0.67	1.00	0.40	0.76	0.06	−18.70	11.72	18.22	0.57
Cold period										
Marousi	PM_{10}	0.91	0.15	1.25	−0.06	−0.16	2.50	9.21	13.47	0.86
Zografou	PM_{10}	0.76	2.21	1.33	−0.52	−0.23	16.79	21.88	45.79	0.86

[a] r is the Pearson correlation coefficient when we used the 25% of the dataset.
[b] NMSE: normalized mean square error.
[c] FA_2: factor of two.
[d] FB: fractional bias.
[e] FV: fractional variance.
[f] MBE: mean bias error.
[g] MAE: mean absolute error.
[h] RMSE: root mean square error.
[i] IA: index of agreement.

Table 6
Statistical model performance measures using an independent dataset, for Helsinki, during the warm period. The letters A and B refer to Cases A and B.

Stations	Pollutant	r[a]	NMSE[b]	FA_2[c]	FB[d]	FV[e]	MBE[f]	MAE[g]	RMSE[h] ($\mu g\ m^{-3}$)	IA[i]
Highest hourly concentrations in a day										
Vallila A	NO_x	0.03	0.42	0.68	0.02	0.65	−1.82	43.62	62.17	0.32
Vallila B	NO_x	0.47	0.37	1.71	−0.37	0.10	44.1	54.08	70.35	0.62
Vallila A	PM_{10}	0.71	0.36	0.39	0.09	0.76	−3.21	12.90	20.74	0.70
Vallila B	PM_{10}	0.73	0.36	0.40	0.11	0.73	−3.9	12.44	20.43	0.71
Kallio A	NO_x	0.22	2.54	0.08	0.7	1.01	−33.19	36.90	69.82	0.36
Kallio B	NO_x	0.34	0.93	0.18	0.05	0.76	−3.00	28.13	59.58	0.42
Kallio A	PM_{10}	0.73	0.13	1.50	−0.16	−0.16	4.66	8.12	9.98	0.82
Kallio B	PM_{10}	0.73	0.13	1.50	−0.16	−0.16	4.66	8.12	9.98	0.82
Daily average concentrations										
Vallila A	PM_{10}	0.81	0.11	1.12	−0.001	0.32	0.012	4.26	6.19	0.88
Vallila B	PM_{10}	0.75	0.16	1.2	−0.003	0.35	0.024	4.64	7.08	0.83
Kallio A	PM_{10}	0.84	0.09	0.78	−0.02	0.15	0.31	3.39	4.51	0.90
Kallio B	PM_{10}	0.84	0.09	0.78	−0.02	0.15	0.31	3.39	4.51	0.90

[a] r is the Pearson correlation coefficient when we used 25% of the dataset.
[b] NMSE: normalized mean square error.
[c] FA_2: factor of two.
[d] FB: fractional bias.
[e] FV: fractional variance.
[f] MBE: mean bias error.
[g] MAE: mean absolute error.
[h] RMSE: root mean square error.
[i] IA: index of agreement.

Another significant variable that could be included in the model is the vehicular traffic flow; however, these measurements are not available for Athens. Local authorities measure traffic volumes in selected roads, but not on a continuous basis. The CO concentrations have therefore previously been used as a proxy variable of the local vehicular traffic (Kassomenos, 2005). It is also possible to use general characteristics of the traffic flows as model input variables (Kassomenos et al., 2006; Vardoulakis and Kassomenos, 2008; Chaloulakou et al., 2003).

The model has some inherent limitations. As the model uses the concentrations and meteorological data measured during the present day as predictor variables, it inherently assumes persistence of both the meteorological and air quality situation. The model therefore cannot allow for temporally rapid changes of the meteorological conditions, or a rapidly changing chemical composition of the air masses. This implies that the model tends to underpredict concentrations during any regional or LRT'ed episodes, such as e.g., those caused by wild-land fires or suspended dust on a regional or larger scale. Rapid transport of air masses could also be caused, for instance, by a land–sea breeze or a recirculation of pollution within an urban area.

We separately checked the performance of the model for PM_{10} in case of land–sea breeze or a recirculation of pollution in Athens. To determine these days we employed the methodology introduced by Kassomenos et al. (1998). Applying this methodology we found that the Pearson correlation coefficient for these days was 0.36, compared with the value of 0.80 for the whole period. Similar low results were found for NO_x. The model is also expected to fail in case of meteorological frontal activity, or in case of other sudden changes in

Table 7
Statistical model performance measures using an independent dataset, for Helsinki, during the cold period. The letters A and B refer to Cases A and B.

Stations	Pollutant	r[a]	NMSE[b]	FA_2[c]	FB[d]	FV[e]	MBE[f]	MAE[g]	RMSE[h] ($\mu g\ m^{-3}$)	IA[i]
Highest hourly concentrations in a day										
Vallila A	NO_x	0.80	1.13	1.8	0.2	0.32	−27.65	79.93	140.18	0.86
Vallila B	NO_x	0.74	1.12	1.59	0.004	0.39	−0.59	93.48	154.36	0.82
Vallila A	PM_{10}	0.60	0.27	0.60	−0.09	0.16	1.33	12.25	17.61	0.77
Vallila B	PM_{10}	0.67	0.26	0.71	−0.02	−0.01	0.33	12.29	17.28	0.81
Kallio A	NO_x	0.64	0.47	1.68	−0.11	0.32	4.05	44.65	58.86	0.77
Kallio B	NO_x	0.58	0.63	1.63	0.08	0.15	−2.8	49.72	65.62	0.75
Kallio A	PM_{10}	0.66	0.67	1.7	−0.01	0.75	0.60	15.67	27.11	0.67
Kallio B	PM_{10}	0.79	0.51	1.9	0.02	0.65	−0.68	13.67	23.22	0.78
Daily average concentrations										
Vallila A	PM_{10}	0.67	0.16	0.57	0.02	0.20	−0.16	6.57	8.43	0.81
Vallila B	PM_{10}	0.69	0.15	0.79	−0.11	0.17	0.97	6.48	8.29	0.82
Kallio A	PM_{10}	0.70	0.13	0.43	−0.03	0.30	0.18	3.94	5.36	0.82
Kallio B	PM_{10}	0.76	0.11	0.68	−0.02	0.26	0.15	3.56	4.93	0.85

[a] r is the Pearson correlation coefficient when we used the 25% of the dataset.
[b] NMSE: normalized mean square error.
[c] FA_2: factor of two.
[d] FB: fractional bias.
[e] FV: fractional variance.
[f] MBE: mean bias error.
[g] MAE: mean absolute error.
[h] RMSE: root mean square error.
[i] IA: index of agreement.

Table 8
Comparison of the results of MLR and ANN methodologies applied to 25% of the dataset for Athens.

Stations	Pollut	MLR (r)	ANN (r)
Marousi/warm period	NO_xHD	0.73	0.85
Marousi/warm period	PM_{10}HD	0.73	0.72
Zwgrafou/warm period	NO_xHD	0.64	0.76
Zwgrafou/warm period	PM_{10}HD	0.59	0.64
Marousi/cold period	NO_xHD	0.69	0.74
Marousi/cold period	PM_{10}HD	0.50	0.61
Zwgrafou/cold period	NO_xHD	0.60	0.62
Zwgrafou/cold period	PM_{10}HD	0.22	0.32
Marousi/warm period	PM_{10}AD	0.81	0.82
Zwgrafou/warm period	PM_{10}AD	0.67	0.77
Marousi/cold period	PM_{10}AD	0.91	0.90
Zwgrafou/cold period	PM_{10}AD	0.76	0.60

HD is the highest daily and AD the average daily air pollutant concentrations, r Person correlation coefficient, when we used 25% of the dataset.

Table 9
Comparison of the results of MLR and ANN methodologies applied to 25% of the dataset for Helsinki.

Stations	Pollut	Case A MLR (r)	Case A ANN (r)	Case B MLR (r)	Case B ANN (r)
Vallila/warm period	NO_xHD	0.03	0.07	0.47	0.56
Vallila/warm period	PM_{10}HD	0.71	0.72	0.73	0.73
Kallio/warm period	NO_xHD	0.22	0.31	0.34	0.44
Kallio/warm period	PM_{10}HD	0.73	0.79	0.73	0.79
Vallila/cold period	NO_xHD	0.80	0.79	0.74	0.79
Vallila/cold period	PM_{10}HD	0.60	0.68	0.67	0.77
Kallio/cold period	NO_xHD	0.64	0.70	0.58	0.66
Kallio/cold period	PM_{10}HD	0.66	0.66	0.79	0.79
Vallila/warm period	PM_{10}AD	0.81	0.86	0.75	0.83
Kallio/warm period	PM_{10}AD	0.84	0.82	0.84	0.91
Vallila/cold period	PM_{10}AD	0.67	0.69	0.69	0.80
Kallio/cold period	PM_{10}AD	0.70	0.77	0.76	0.81

HD is the highest daily and AD the average daily air pollutant concentrations. r is the Pearson correlation coefficient, when we used 25% of the dataset. The letters A and B refer to cases A and B (defined in the text).

the weather conditions. Clearly, the accuracy of the model is also crucially dependent on the selection and preparation of the input variables. For instance, it was found out that the use of the meteorologically pre-processed data significantly improved the accuracy of the model forecasts. However, such a pre-processed meteorological data was not available in case of Athens.

To test the prognostic ability of the MLR model with other existing models, we have compared the results obtained by MLR with those by the ANN model described in Section 2.3. Comparing the two evaluations made using the two sample datasets (consisting of the 25% of the total data), we found out that the ANN model provides slightly better results. Table 8 presents the results for Athens and Table 9 the comparisons for Helsinki (Cases A and B). As we could see in Tables 8 and 9, the Pearson correlation coefficient (r) for the sample of 25% of the entire dataset is slightly better than the relevant parameter computed by the MLR.

In previous research, several authors (Comrie, 1997; Kukkonen et al., 2003, among others) have compared various alternative methodologies (including linear statistical models, ANN models and deterministic models) in different environments. For instance, Kukkonen et al. (2003) found out that the ANN models used provided better results, compared with simpler linear statistical techniques. Although in the present study it was found that the ANN model behaves slightly better than MLR, this simpler statistical model has other advantages. First, the MLR model is substantially simpler to construct and use. Second, the functioning of the model and the results can be interpreted in a more straightforward way. In particular, the user of MLR can know in detail, which is the exact contribution of each predictor to the final result; this is commonly not known in case of ANN models. The use of simpler statistical techniques can therefore in some cases be preferable to the use of ANN models.

3.4. Time series of the measured and predicted concentrations

Time series of the measured and predicted concentrations for both cities are presented in Figs. 4–7. Fig. 4 a–b presents the highest hourly concentrations of NO_x and PM_{10} for Marousi station during the warm period, while Fig. 5 a–b indicates the daily average concentrations of PM_{10} in Marousi station, for both periods. Fig. 6 a–b indicates the highest hourly concentration of NO_x and PM_{10} in Vallila station (Case A) and Kallio station (Case B) during the cold period. In the case of daily average concentration of PM_{10}, Fig. 7 a–b illustrates the observed and predicted values at the stations of Vallila (Case A) in warm period and Kallio station (Case B), in cold period respectively.

The temporal variation of the curves is similar for most of the predicted and observed values. However, there is substantial under- or overprediction in some cases. In case of the highest hourly concentrations of NO_x and PM_{10} (Figs. 4–6), the model seems on the average to slightly overpredict these in both cities, while for the daily average concentrations of PM_{10} the model seems on the average to slightly underpredict (Fig. 7 a–b).

4. Conclusions

This study describes a linear model to forecast the highest hourly concentrations of NO_x and PM_{10}, as well as the daily average

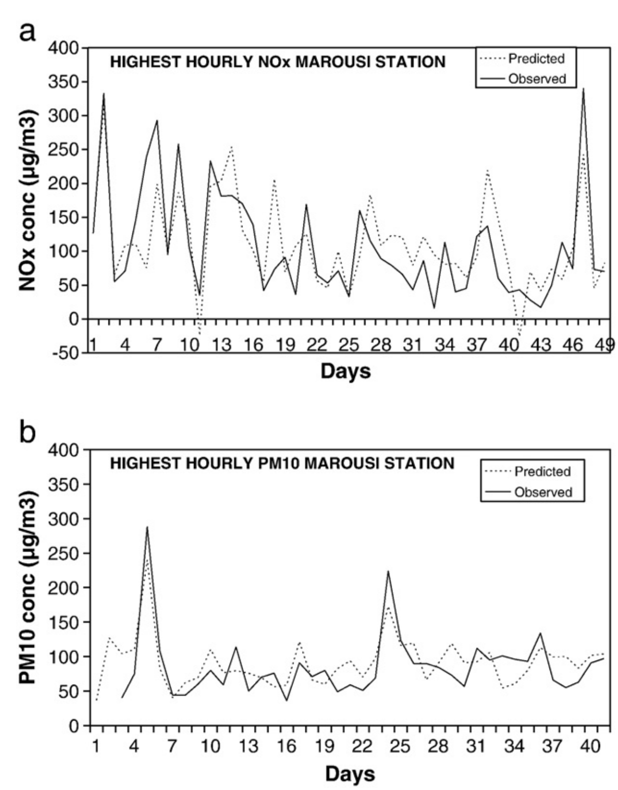

Fig. 4. Highest hourly concentrations of (a) NO_x and (b) PM_{10} of observed and predicted values, at Marousi, during warm period (days are a random collection of data from 25% of the initial dataset covering the period of 6 months).

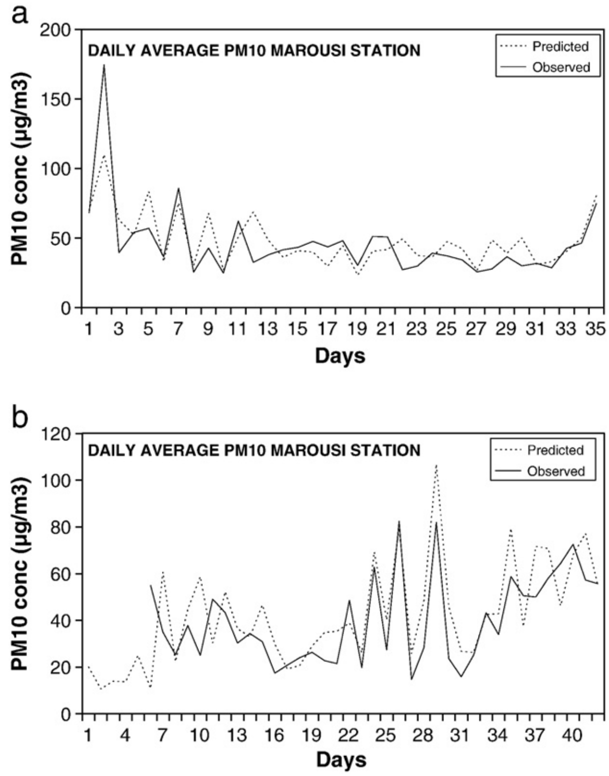

Fig. 5. Daily average concentrations of PM_{10} of (a) observed and (b) predicted values, at Marousi, during warm period and cold period, respectively (days are a random collection of data from 25% of the initial dataset covering the period of 6 months).

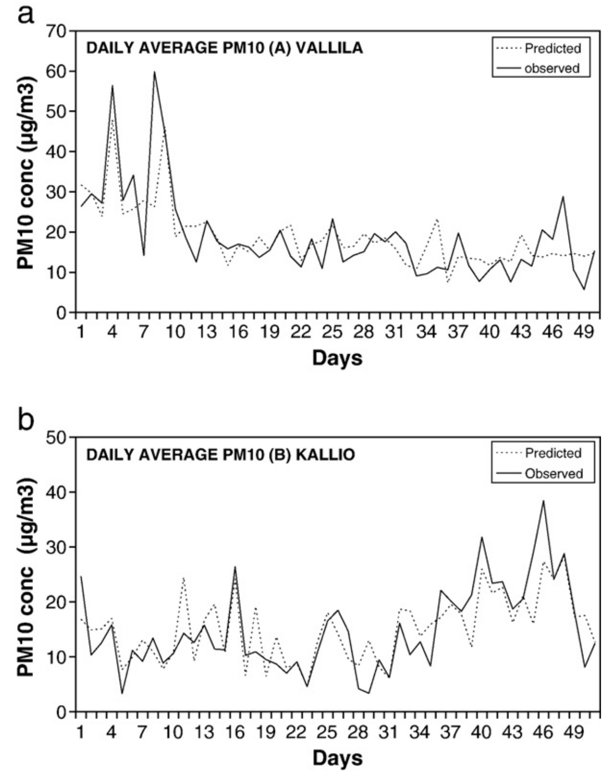

Fig. 7. Daily average concentrations of PM_{10} of (a) observed and (b) predicted values, at Vallila (Case A) and Kallio (Case B), during warm and cold periods respectively (days are a random collection of data from 25% of the initial dataset covering the period of 6 months).

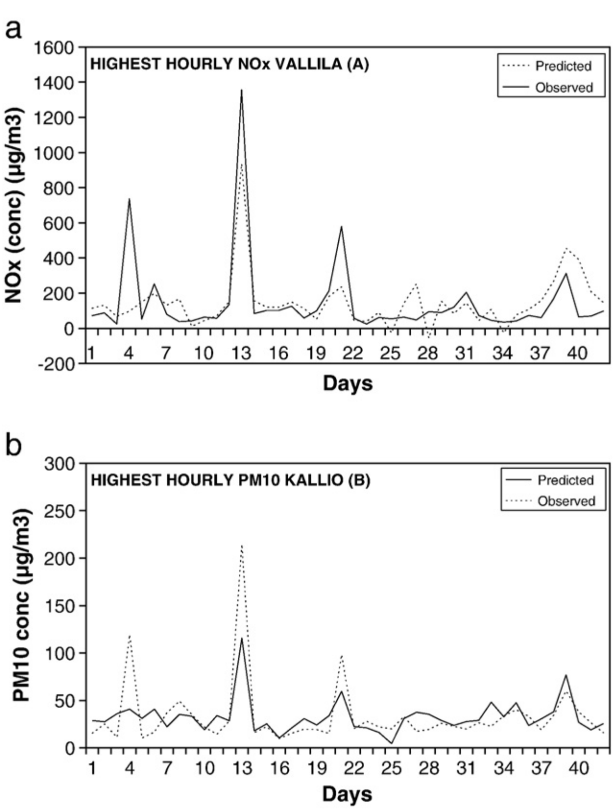

Fig. 6. Highest hourly concentrations of (a) NO_x and (b) PM_{10} of observed and predicted values, at Vallila (Case A) and Kallio (Case B) respectively, during cold (days are a random collection of data from 25% of the initial dataset covering the period of 6 months).

concentrations of PM_{10} of the next day, using air quality and meteorological data. The model was applied and evaluated against the data measured in 2005, for Athens and Helsinki. We have chosen two stations for each city, representative of urban background and urban traffic. The dataset was separated in two seasons, the warm (from April to September) and cold (from October to March) period. Finally, we use a variety of evaluation methods in order to check the performance of the models. Model evaluation statistics and prediction skills were computed against an independent dataset that contains 25% of the total dataset.

In Athens and Helsinki, there are substantially different types of climate, different terrain and topographic features, and differing amount of population and population density. As expected, both the measured NO_x and PM_{10} concentrations were substantially higher in Athens, compared with Helsinki.

Generally, the best predictors in both cities were the concentrations of NO_x and PM_{10} during the evening hours of the present day. The variables of meteorological data had expectedly a substantial contribution in both cities. In Athens, where only locally measured meteorological values were used as model input variables, the wind speed and wind direction index were important variables. In Helsinki, where also the meteorologically pre-processed variables were used, the most significant meteorological variables were the inverse Monin-Obukhov length, and the mixing height. This is understandable, as these parameters describe physically the most important characteristics of atmospheric diffusion and turbulence conditions.

It is interesting to note that the local wind regimes (i.e., the urban wind speed and direction patterns) in Athens substantially affect the dispersion and accumulation of the released air pollutants. Such wind regimes and the related atmospheric stability conditions are to a large extent formed by the specific topography and land-use characteristics of the area. Athens is located in a relatively narrow densely populated basin, surrounded by high mountains on three sides and the sea on one side of the basin. It is therefore to be expected that wind patterns have a significant contribution on the formation of air pollution levels in Athens.

Such a special topography does not exist in case of Helsinki, which is situated in a fairly flat coastal region. Especially in northern European conditions, local emissions can cause air pollution episodes in case of extremely stable stability conditions, under prevailing strong ground-based or low-level inversions in winter (e.g., Kukkonen et al., 2005). It was therefore understandable that wind speed was not found to be a significant predictor variable in Helsinki; instead the Monin-Obukhov length and the mixing height were found to be more important.

For both cities, the model expectedly performed better for the forecasting of the daily average concentrations, compared with the highest hourly concentrations of the next day. For Helsinki, we executed the model both without and with the Monin-Obukhov length and the mixing height. The model expectedly had a better performance, when the above mentioned turbulence and diffusion parameters were included.

The model had a higher prediction accuracy in Helsinki than Athens. The main reasons for this result are probably associated with the more complex structure of the emission sources and the more complex topography in the surroundings of Athens. The latter may result in complicated meteorological phenomena, such as commonly occurring land–sea breeze and recirculation of pollution. The factors related to the higher population and the higher population density, combined with specific socioeconomic conditions probably also cause uncertainties to the predictions (Kassomenos, 2005).

The limitations of the model include that it inherently assumes persistence of the both the meteorological and air quality situation. The model therefore cannot allow for rapid temporal changes (of the order of less than one day) caused by regional or LRT'ed episodes, such as e.g., those caused by wild land fires or suspended dust on a larger scale. The model is also expected to be inaccurate in case of sudden changes in the weather conditions, such as, e.g., those associated with meteorological frontal activity.

Acknowledgements

The research leading to these results has been part of the COST ES0602 action “Towards a European Network on Chemical Weather Forecasting and Information System” and has received funding from the European Union's Seventh Framework Programme FP/2007–2011 under grant agreement no. 212520 (MEGAPOLI) and the project “Transport related Air Pollution and Health impacts—Integrated Methodologies for Assessing Particulate Matter” (TRANSPHORM).

References

Aarnio P, Martikainen J, Hussein T, Valkama I, Vehkamaki T, Sogacheva L, et al. Analysis and evaluation of selected PM_{10} pollution episodes in the Helsinki Metropolitan area in 2002. Atmos Environ 2007;42(17):3992–4005.

Abraham J, Comrie C. Real-time ozone mapping using a regression-interpolation hybrid approach, applied to Tucson, Arizona. J Air Waste Manage Assoc 2004;54:914–25.

Athanassiadou M, Flocas H, Harrison MA, Hort MC, Witham CS, Millington S. The dust event of 17 April 2005 over Athens, Greece. Weather 2006;61(5):125–31.

Barcenas O, Olivas E, Guerrero JD, Valls G, Rodriguez C, Tascon S. Unbiased sensitivity analysis and pruning techniques in neural networks for surface ozone modeling. Ecol Model 2005;182:149–58.

Basurko E, Berastegi G, Madariaga I. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environ Model Softw 2006;21:430–46.

Borge R, Lumbreras J, Vardoulakis S, Kassomenos P, Rodriguez E. Analysis of long range transport influences on urban PM_{10} using two stage clusters. Atmos Environ 2007;41(21):4434–50.

Chaloulakou A, Grivas G, Spyrellis N. Neural Network and multiple regression models for PM_{10} prediction in Athens: a comparative assessment. Air Waste management Assoc 2003;53:1183–90.

Chaloulakou A, Kassomenos P, Grivas G, Spyrellis N. Continuous measurements of $PM_{10}/PM_{2.5}$ and black smoke concentrations, at a fixed site in Athens, Greece. Environ Int 2005;31:651–9.

Comrie AC. Comparing neural networks and regression models for ozone forecasting. J Air Waste Manage Assoc 1997;47:653–63.

Comrie A, Diem J. Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona. Atmos Environ 1999;33:5023–36.

Cordelino C, Chang M, St John J, Murphey B, Cordle J, Ballagas R, et al. Ozone prediction in Atlanta Georgia: analysis of the 1999 ozone season. J Air Waste Manage Assoc 2001;51:1227–36.

Dennis JE, Schnabel RB. Numerical methods for Unconstrained Optimization and Nonlinear equations. Englewood Cliffs, NJ: Prentice-Hall; 1983.

Dockery DW, Pope III CA. Acute respiratory effects of particulate air pollution. Annu Rev Public Health 1994;15:107–32.

Gardner MW, Dorling SR. Statistical surface ozone models: an improved methodology to account for non linear behaviour. Atmos Environ 2000;34:21–34.

Karppinen A, Kukkonen J, Elolahde T, Konttinen M, Koskentalo T, Rantakrans E. A modeling system for predicting urban air pollution: model description and applications in the Helsinki metropolitan area. Atmos Environ 2000a;34:3723–33.

Karppinen A, Joffre SM, Kukkonen J. The refinement of a meteorological preprocessor for the urban environment. Int J Environ Pollut 2000b;14(1–6):565–72.

Kassomenos P. Socioeconomic aspects in an extended contemporary city—how can we approach them by using a pollutant indicator 2004. Water Air Soil Pollut 2005;162:315–29.

Kassomenos P, Flocas HA, Lykoudis S, Skouloudis A. Spatial and temporal characteristics of the relationship between air quality status and mesoscale circulation over an urban Mediterranean basin. Sci Total Environ 1998;217:37–57.

Kassomenos P, Lykoudis S, Kallos G. Winter southern wind flow and air pollution episodes over Athens, Greece. Glob Nest Int J 1999:99-110.

Kassomenos P, Karakitsios S, Papaloukas C. Estimation of the daily traffic emissions in a south-European urban complex during a workday. Evaluation of several “what if” scenarios. Sci Total Environ 2006;370:480–90.

Kassomenos P, Papaloukas C, Petrakis M, Karakitsios S. Assessment and prediction of short term hospital admissions. The case of Athens, Greece. Atmos Environ 2008;42:7078–86.

Kolehmainen M, Martikainen H, Ruuskanen J. Neural networks and periodic components used in air quality forecasting. Atmos Environ 2001;35:815–25.

Kukkonen J, Partanen L, Karpinnen A, Ruuskanen J, Junninen H, Kolehmainen M, et al. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modeling system and measurements in central Helsinki. Atmos Environ 2003;37:4539–50.

Kukkonen J, Pohjola M, Sokhi RS, Luhana L, Kitwiroon N, Frangou L, et al. Analysis and evaluation of selected local-scale PM_{10} air pollution episodes in four European cities: Helsinki, London, Milan, Oslo. Atmos Environ 2005;39:2759–73.

Lalas D, Veirs VR, Karras G, Kallos G. An analysis of the SO_2 concentration in Athens, Greece. Atmos Environ 1982;16(3):531–44.

Liu Yang, Kahn R, Chaloulakou A, Koutrakis P. Analysis of the impact of forest fires in August 2007 on air quality of Athens using multi-sensor aerosol remote sensing data. meteorology and surface observations, 43. ; 2009. p. 3310–8.

Lu H-C. The statistical characters of PM_{10} concentration in Taiwan area. Atmos Environ 2003;36:491–502.

Lykoudis S, Psounis N, Mavrakis A, Christides A. Predicting photochemical pollution in an industrial area. Environ Monit Assess 2008;142(1–3):279–88.

Niska H, Hiltunen T, Karppinen A, Ruuskanen J, Kolehmainen M. Evolving the neural network model for forecasting air pollution time series. Eng Applic Of Art Intell 2004;17:159–67.

Niska H, Rantamäki M, Hiltunen T, Karppinen A, Kukkonen J, Ruuskanen J, et al. Evaluation of an integrated modeling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. Atmos Environ 2005;39:6524–36.

Olesen HR. The model validation exercise at Mol: overview of results, workshop on Operational short-range atmospheric dispersion models for environmental impact assessment in Europe. Int J Environ Pollut 1995;5:761–84.

Ordieres JB, Vergara E, Capuz R, Salazar R. Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juarez (Chihuahua). Environ Model Softw 2004;20:547–59.

Paschalidou A, Kassomenos P. Comparison of air pollutant concentrations between weekdays and weekends in Athens, Greece for various meteorological conditions. Environ Technol 2004;25(11):1241–55.

Paschalidou K, Kassomenos P, Bartzokas A. A comparative study on various statistical techniques predicting ozone concentrations: implications to environmental management. Environ Monit Assess 2009;148:277–89.

Ping Shi JI, Harrison R. Regression modeling of hourly NO_x and NO_2 concentrations in urban air in London. Atmos Environ 1997;31:4081–94.

Sofiev M, Vankevich R, Lotjonen M, Prank M, Petukhov V, Ermakova T, et al. An operational system for the assimilation of satellite information on wild-land fires for the needs of air quality modelling and forecasting. Atmos Chem Phys Discuss 2009;9:6483–513http://www.atmos-chem-phys-discuss.net/papers_in_open_discussion.html.

SPSS Inc. Advanced techniques: regression (SPSS 10.0). SPSS Inc.; 2000.

Vardoulakis S, Kassomenos P. Comparison of factors influencing PM_{10} levels in Athens (Greece) and Birmingham (UK). Atmos Environ 2008;42(17):3949–63.

Willmott CJ. On the validation of models. Phys Geogr 1981;2:184–94.

Ziomas I, Melas D, Zerefos C, Bais A. Forecasting peak pollutants levels from meteorological variables. Atmos Environ 1995;29:3703–11.